CSC 226

WEB SEARCH AND INFORMATION RETRIEVAL (2units E)

1. The Concept of Information Retrieval

The phrase "information retrieval" was first coined in 1952, according to Spack and Willet (1997), and it acquired prominence in the research community in 1961. At the time, the organizational role of information retrieval was considered as a significant advancement in libraries, which were no longer merely book stores, but also locations where information was cataloged and indexed. The idea behind information retrieval is that there are documents or records holding data that have been structured in a way that makes it easy to discover. As a result, an information retrieval system is implemented to obtain the documents or information that the user community requires. It should provide the appropriate information to the appropriate user.

1.1 Definition of Information Retrieval (IR)

The effectiveness of a library is determined by the speed with which its users can locate and retrieve accurate and relevant information for their multifaceted needs. Chimah, **Unagha and Nwokocha (2010)** defined information retrieval as “a process that involves extraction of information from a collection or database in response to an information problem”. **Edom (2012)** averred that information retrieval is “a mechanism or apparatus that aids library users to locate, retrieve and utilize needed documents, information or books from the library collection”. **Rashid (2020)** described information retrieval as “the activity of obtaining the right information, to the right user at the right time and that it deals mainly with the representation, storage, organization and access to information”. It is also the activity of obtaining information resources relevant for a user’s information need from a collection of information resources (Manning, **Raghavan and Schutize, 2009)**. **Chowdhury (2017)** opined that information retrieval is “concerned with all activities related to the organization of, processing of, and access to, information of all forms and formats. That it allows people to communicate with an information system or service in order to find information – text, graphic images and sound recordings or video that meets their specific needs”.

1.2 Elements or Components of Information Retrieval

Information Retrieval mainly consists of four elements which are briefly discussed below:

Information Carrier: This serves the purpose of transmitting information from one system to another. It is also storage device used to store information for future use. Examples are magnetic tapes, CD-ROM, DVD, and Memory stick, etc.

Descriptor: This is a term or terminology that is used to search for information from storage known as a descriptor. It can be a name, number, signal, etc.

Document address: Every document must have an author that identifies the location of those documents. In Library and Information Science profession, document address involves a call number, class number, ISBN, shelf number, accession number, file number, etc. all these help users in retrieving their required information.

Transmission of Information: Transmission of information means to supply any document that is required by the user. Information retrieval system uses various communication channels to do this. It works like source → transmitter → media →receiver → destination.

1.**3 Functions of Information Retrieval**

These include the following:

i. **Acquisition**: Acquisition is the first function of information retrieval. It is the process of selecting, ordering and receiving materials for library or archival collections. It can be through purchase, donation or gift and the resource can be books, documents, journals, etc.

ii. **Content Analysis:** The second step is to evaluate the effectiveness of the information acquired.

iii. **Content Presentation:** Information should be presented clearly and effectively so that users will be able to understand them easily. For this purpose (catalogue, bibliography, index, CAS) will help a lot.

iv. **Creation of Store:** In this stage the library authority creates a new file for storing their information already collected and ready for presentation. Such information is always organized in a systematic way.

v. **Creation of Search Method:** The authority will decide what kind of search logic they may use for searching and retrieving information.

vi. **Dissemination**: This is the last stage of information retrieval. It is an act of spreading information systematically. So the whole function happens like Acquisition → Content analysis → Information presentation →Creation of store → Creation of search logic →Dissemination retrieval result → Stop.

1.4 Techniques of Information Retrieval

Two techniques are used to retrieve information effectively and they are briefly described in below:

1. **Traditional System:** Traditional system involves the following segments:

i. **Catalogue:** A catalogue is a list of books that libraries use to locate items. It includes information such as the author's name, title, edition, number of volumes, publication date, page number, series name, and ISBN. It is a necessary tool for information retrieval.

ii. **Index**: An index is a list of entries organized in a systematic way to help library patrons find information in a document. It is an important tool for retrieving information.

iii. **Abstract**: This is an exact and succinct representation of the original material. It allows users to get a general idea of the content without having to read it completely. It can also be used to retrieve information.

iv. **Bibliography**: A bibliography is a list of books that is not restricted to a single collection.

v. **Authority File:** Library Authority will sometimes use its own methods to retrieve information rather than using current methods. This strategy is frequently written down in books or files for future use.

2. **Non-traditional System:** Non-traditional systems are divided into two main categories namely:

i. **Semi-automatic System**: A semi-automatic system combines human and machine retrieval of data. This method necessitates human intelligence and labor in addition to the usage of machines. Semi-automatic systems include catalog cards, punch cards, notes cards, apache cards, and so on.

ii. **Automated System**: This system is mostly dependent on a single computer and networking technology. In the 1970s, an offline network was used, but from 1980 to the present, real-time sharing has been used. Computer, modem, CD-ROM, hard disk, internet, and other automatic tools are examples. The following are their descriptions:

a) Computer: An electronic machine that can store and work with large amounts of information.

b) CD-ROM: A compact disc used as a read-only optical memory device for a computer system.

c) Hard Disk: A right non-removable magnetic disk with a large data storage capacity.

d) Floppy Disk: A flexible removable magnetic disk (typically encased in a hard-plastic shelf) for storing data.

e) The Internet: A global computer network made up of interconnected networks that use standardized communication protocols to provide a variety of information and communication services.

Conclusion

Information retrieval focuses on the provision of a reliable mechanism that facilitates speedy access to required information from the total collection with highest level of accuracy and precision. It guarantees user satisfaction and optimal utilization of information resources. The effectiveness of libraries as a reservoir of knowledge and the memory of society lies on the functionality and capabilities of the information retrieval framework. The usefulness of acquisition, organization through classification and cataloging as well as storage of information resources can only be appreciated if users can retrieve the material with ease as and when needed. Information retrieval is pervasive and thus applicable to all formats of information.

2. Objectives of Information Retrieval

The core objective of information retrieval is to ferret out relevant information or information surrogates as and when needed from the mass of collection or database in response to a user's query / request. Onwuchekwa (2011) highlighted eight objectives an information retrieval must seek to actualize as follows:

1. Prompt dissemination of information;
2. Filtering of information;
3. The right amount of information at the right time;
4. Browsing capability;
5. Getting information in an economical way;
6. Provision of current literature;
7. Interpersonal communication; and
8. Personal help.

1. **Prompt Dissemination of Information:** This is predicated on the information retrieval system meeting set goals within a timely frame. The response rate to the user's enquiry must be at such speed that guarantees effectiveness. Any form of sluggishness or downtime tends to frustrate the retrieval effort. The system is expected to execute commands without delay and perform regularly within the specified time limit.

2. **Filtering of Information:** This involves removal of unwanted information from the maze of information using digital mechanism. It seeks to tailor retrieval to the preferences of the user and search specification. Every good retrieval system must therefore have the capability of sorting and discriminating against or restricting the flow of undesirable information. It selects the most valuable to meet the user's search queries.

3. **The right amount of information at the right time**: Corollary to the issue of filtering is that the system should provide the right amount of information at the right time to avoid information overload. Excessive amount of information could be overwhelming and makes comprehension as well as utilization pretty difficult. Since retrieval of too much information tends to be counterproductive, the objective of the system to ensure that the right amount information is retrieved at the right time.

4. **Browsing capability:** It provides an embedded application or interface which enables users to locate, view and retrieve relevant information. There is usually a dialogue box through which

search queries are sent to the system. It crawls through the databases to ferret out information that matches the search terms and the output is displayed as feedback to the patron's queries. Browsing capability is therefore a core characteristic of any retrieval system. The user needs to understand the peculiarities of each system's browsing capability and features for optimal utilization.

5. **Getting information in an economical way:** Some of the barriers to information are cost, language and distance. If users pay prohibitive fees to retrieve information or pay to translate the information into intelligible language or format, it is therefore not economical. All forms of barriers that make it expensive to retrieve appropriate information must be eliminated for effectiveness. This is particularly important in provision of information to the visually impaired persons and other disadvantaged persons in society.

6. **Provision of current literature:** Information retrieval systems should enhance retrieval of up-to-date information in any field of choice. This helps to ensure reliability and authenticity of contents. It should also have the capability to enable users delineate the scope or coverage or period desired for retrieval. For example, the user can request that only information covering the period of 2019-2021 alone should be retrieved. It stands to reason that the system must be constantly up-to-date with latest information or literature to maintain relevance as users abhor obsolete or stale information.

7. **Interpersonal communication:** Modern retrieval systems provide for interactive interface. This allows for online chats, newsgroups, discussion forums, video conferencing and email. Online chat which is the most common involves users communicating using the keyboard and receiving feedback as well. It could be synchronous or asynchronous. The former depicts simultaneous and instant communication. The latter involves leaving a message and receiving feedback later.

8. **Personal help:** Personal help features are integral part of information retrieval system. It provides an interface for users having difficulties in navigating the system to request for specific helps or services. It makes available detailed procedures to resolve retrieval problems or trouble-shooting guidance. This helps to minimize or eliminate user frustrations in the process of searching for requisite information.

2. **Information Representation and retrieval**

Chu-Heting (2003) defined information representation as “” the extraction of some elements (keywords or phrases) from a document or the assignment of terms (descriptors or subject headings) to a document so that its essence can be characterized and presented.

Information representation can be done via any combination of the following means: abstracting, indexing, categorization, summarization and extraction”. Roshdi and Roohparvar (2015) and Djoerd (2009) noted that information representation is called the indexing process which is done behind the scene and excludes the involvement of end-users. It is chiefly aimed at the creation of terms, words, notations, tags that significantly and sufficiently characterize the information resources in the collection for uniqueness thereby facilitating retrieve ability.

2.1 Approaches to Information Retrieval

Chu-Heting (2003) identified the different approaches for information representation in library service as follows:

- Alphabetical subject approach
- Hierarchical or Subject Indexing approach
- Coordinate Representation approach
- Computer-based application approach
- Statistical Methods approach

a. Alphabetical Subject Approach

This involves the grouping of information and surrogates under subjects and further arranged into alphabetic order for easy retrieval. The library professional must clearly resolve the issue of synonyms, homographs, singular or plural forms, complex and compound words or subjects in the alphabetical subject approach. It also recognizes and caters for relationships among subjects and terms; these include syntactic and semantic relationships. The former refers to relationships among words and phrases as depicted by their arrangement. For example, a keyword search for “photographs and Albums” should permit patrons to indicate if they want “photographs of Albums” or “Albums of photographs”. The latter related with the meanings of words. For example, there is semantic difference between mercury (planet) and mercury (metal) irrespective of similarity in sound and spelling.

b. Hierarchical or Subject Indexing Approach

This is one of the traditional approaches which rely on subject and classification schemes to provide content or information representation for documents and other materials. It is the mechanism of identifying and selecting descriptors, taxonomic categories, notations or terms which adequately express the contents of the information resources. This helps to facilitate the retrieval of information based on its subject content.

c. Computer-based Application Approach

This approach deals with the deployment of computer applications or software for indexing or recording information. It is a departure from manual subject indexing to automatic indexing and abstracting. The application collocates frequently occurring keywords to represent plural facets of a document. Two notable computer applications in this regard are keyword In Context (KWIC) and Keyword out of Context (KWOC). The former identifies high-frequency keywords and the latter; proximity of non- significant keywords outside the document. The combinations of keywords in context and out of context are used for constructing appropriate information representation.

d. Statistical Methods

This is advancement over the computer – based application approach. It leverages on Machine Readable Cataloging (MARC) standards and other algorithms to represent internet resources and other global networked information resources. The use of transmission control protocol and the Internet protocol (TCP/IP) help to mainstream MARC records of library collection (sounds, videos, audios, images and other multimedia objects) into the cyberspace, thus library collections which had been converted into MARC records can be virtually represented on the web.

2.2 Language in Information Retrieval

An indexing language is a set of terms or codes that can be used as an index's access point. In contrast, a searching language refers to the terms that a searcher uses to express a search criteria. During a search, the same terms or codes can be used as access points to records. The ultimate objective of information retrieval is locating accurate and relevant information with ease and within the quickest timeframe to meet the patron’s need. This is only made possible when the language of searching matches the indexing language used in information representation. Indexing language is the fulcrum upon which information representation and retrieval thrive.

There are basically three types of indexing languages, to wit:

(a) Natural Indexing Language
(b)Free Indexing Language
(c) Controlled Indexing Language
Each of these is briefly discussed below.

a. Natural Indexing Languages: Natural Indexing Languages or Derived Term Systems: Natural indexing languages or derived term systems are not really a distinct or ordinary language of the document being indexed, rather than a natural or ordinary language. To put it another way, this is not artificial language but language used in normal communication or ordinary context of the document. This means that it is derived from terms and phrases contained in the document. A derived term system is one in which all descriptors are derived from the indexed document. It is premised that the terms or phrases used are discipline specific or terminologies that are familiar to the subject hence not strange to users for information retrieval. Natural indexing language could be created manually or automatically by computer indexing. The computer may index every phrase in the text, with the exception of a restricted stop-list of all common terms, or it may only index terms from a computer-held thesaurus. That is to say the indexer or computer assigns the descriptor derived from the title, abstracts or the full-text of the document. Indexing using natural language is not subject to restricted vocabulary and enjoys wide latitude of terms.

b. Free Indexing Language: Free indexing language differs from natural indexing language in that the latter is limited by the language of the document being indexed, whilst the former is not (i.e., any appropriate term can be assigned). Unlike natural indexing language that is restricted by the terms and phrases contained in the document, free indexing language draws from anywhere adjudged suitable for representing the contents of the document. This means that free language indexing is dependent on the prerogative and skills of the indexer. Furthermore, free language indexing might be done manually by a human indexer, with the quality of the index relying heavily on the indexer's understanding of the subject and terminology.

Because the computer must have some basis on which to assign terms, computerized free indexing is, for all intents and purposes, the same as natural-language indexing.. Terms found adequate and relevant are assigned as index terms.

c. **Controlled Indexing Language:** This is an artificial language that is clearly defined and provides prescribed terms and notation for indexing of documents and retrieval. Many information retrieval applications make extensive use of natural language indexing and controlled language indexing. It provides array of terms as authority list for selecting approved terms for representing documents. This system allows for standardization and consistency among indexers as the choice of subject descriptors must be in tandem with the prescribed guidelines. It involves assigning descriptors that match the concepts of the document as specified from the authority list. Controlled indexing languages are said to be more consistent, making searchers' jobs easier and more efficient. Controlled language indexing, on the other hand, is considered as beneficial in a supporting environment for experienced users since it eliminates the need to negotiate all of the differences inherent in natural language.

3. Information Retrieval Techniques and Models

Today, information can be found in the different media and in different form. Aside storing information in printed document, publishers of information now store information in non-print media like electronic databases, Meta-data, and the internet. The majority of web sites include semi-structured and dynamic material that is intertwined with links and difficult to obtain.

3.1 Information Retrieval Techniques

Scientists and other scholars have established different information retrieval techniques. Among those techniques is the Boolean search technique, term truncation, stemming, amongst others.

a. Boolean Searching

The means of specifying a combination of keywords that must be matched for successful retrieval is known as search logic. As a result, most systems search using Boolean search logic.

Boolean searching is the search technique used in querying the internet or a database. The logic is used to connect terms that describe the concepts in the search query. The technique involves combining keywords or phrases in a single search query to retrieve on relevant and desired search results. It can connect terms from both controlled and natural indexing languages, and both. In order to frame a search statement, search logic may link up to 30 or more search phrases together.

The Boolean technique makes use of the Boolean logic or operators to refine or limit or widen search results. The Boolean logic operators are: AND, OR, and NOT. By applying these operators 'AND', 'OR' and 'NOT', the researcher is able to teach the system the information to include or leave out in a search results. This saves the researcher's time and allows him/her to focus on only the relevant information. The Boolean search technique is often applied when one need more than one word to describe a search problem.

The 'AND' operator is used when one needs to retrieve all information about two or more keywords. For instance, a search term containing the words 'Mosquito and Malaria' will bring results containing both Mosquito and Malaria. The use of the 'AND' operator allows the researcher to combine more than one search term in a single query, which in this case could be Malaria 'AND' Mosquito.

The 'AND' operator could be used if a researcher wants to retrieved search results related to celebrities with sport cars. The researcher could use the search term celebrities 'AND' sport cars. This will fetch out relevant results related to celebrities having dealings with sport cars.

The 'OR' operator is used to retrieve search results containing any of the search terms. This allows this researcher to still combine search terms in a single query and retrieve results on any of the search terms. In other words, the 'OR' operator allows the researcher to make a broader search as such search could yield more results. For instance, a researcher seeking to retrieve information on transmittable viruses could search for Corona virus or Human papilloma virus**.**

The 'NOT' operator is used to narrow the results for a single query. The use of 'Not' helps the researcher to eliminate or exclude (irrelevant) terms or records in a research result. For instance, a researcher interested in searching for information on the 2019 Corona virus could use a search term as "SARS-CoV-2 NOT MERS-CoV". The search will focus on only one type of Corona virus, which is SARS-CoV-2.A researcher interested in reading about Gombe as city in Gombe state could use the search term: `Gombe ‘NOT’ State` to narrow down his or her search results. This will eliminate results about Gombe as a state.

b. Stemming

Stemming involves the practice of producing a morphological variant of a root word. It is generally a language dependent procedure involved in the removal of suffixes from words so that words with the same root match each other. For instance, words such as ‘likes’, ‘likely’, ‘liked’ and ‘liking’ could be stemmed to the root word ‘like’. Stemming allows the researcher to reduce words to its root without significantly affecting the meaning or purpose of the search.

c. Algorism

Algorism in information retrieval is concerned with the way data are filtered, indexed and retrieved with high level of precision in a single query. Consequently, information retrieval algorithm can be classified into three main classes. They are retrieval, indexing, and filtering algorithm. The retrieval algorithm, being the main class of algorithm in information retrieval, focus on how information are extracted from a word-based databank.

3.2. Precision and Recall

Precision and recall are two measuring technique in the sphere of information retrieval used to measure how well an information retrieval system retrieves documents that are relevant to the researcher. Typically, in an information system, the total data base and the information to be retrieved could be divided in four main categories. They include relevant and retrieved; relevant and not retrieved; non relevant and retrieved and non-relevant and not retrieved. While the relevant information helps the researcher to answer his/her questions, the non-relevant information does not contribute to the solving of the research problem. Consequently, in a single query, there is the possibility of retrieving relevant information and also the possibility of retrieving non relevant information. Therefore, precision is the total number of relevant retrieved documents (information useful to the researcher) over the total number of retrieved documents from a single query. Hence, precision encompasses the ability to retrieve top ranked relevant documents in a single search. For simplicity sake, the description of precision could be given as:

$$\text{Precision} = \frac{\text{Number of Retrieved Relevant documents}}{\text{Number of Total Retrieved documents}}$$

On the other hand, recall is the ratio of retrieved relevant document to the number of possible relevant documents in the database. Basically, recall measures the extent to which a retrieval system is able to retrieve all relevant documents in a database. The formula for recall is given as:

$$\text{Recall} = \frac{\text{Number of Retrieved Relevant documents}}{\text{Number of Possible Relevant documents}}$$

3.3 Functional Process of Information Retrieval

Information retrieval comprises some functional processes. While several opinions and thoughts have been raised in the past regarding the functional process of information retrieval, it is now widely accepted that the functional process of information retrieval is composed of four main parts namely; item normalization, selective dissemination of information (i.e. mail), archival document database search, and index database search. These four parts describe the functional process of information storage and retrieval.

i. Item Normalization

In the functional process of information retrieval, item normalization is usually considered the first step in the process since it is at this stage that the format of an incoming item is being normalized in an integrated system. In this stage, the system is able to normalized all external item into a conformable system and also logically restructure all items in the system. The normalization of data here provide room for the creation of searchable data structure as well as stemming and characterization of tokens. This ensures standardization of all items or records in the system as the system could have single formats for all items or allow multiple formats.

ii. Selective Dissemination of Information

This is one of the functional processes of information retrieval characterized by search process, indication of interest by users and the user mail files. At this stage, one is able to match each user need or interest to given or up-coming information in the system and when the user needs is matched with the user, the information is then sent to the user via mail. The information system is able to track a user profile and understand their search interest through their search history and then provide future recommendations to the users in the light of the accumulated information.

While this process is primarily used in a text based environment, it can also be used in a multimedia environment.

iii. Document Database Search

The document database search primarily comprises three segments namely; the document database that comprises all information entered and stored in the database, the search query or search term entered by the researcher, and the search process. At this stage, a query entered by the user search through the database to retrieve relevant information based on the search query entered. This search is mainly considered retrospective and the user may benefit from the selective dissemination of information feature if he/she is online. Users may be able to filter their search at this stage depending on the interface and feature of the design database.

iv. Index Database Search

This process involves searching through an index database or a collection indexes to retrieve an earlier saved or stored file for planning or decision making. In this search, information retrieved only direct the user to where he/she can acquire detailed information. The information in the database only point one to where detailed information could retrieve. The index database houses multiple index files. Index files are classified into two namely; public and private index files. The classifications of these files are usually set at the point of entering the data in the database.

Metadata-based search is the second option. Metadata-based search uses meta-descriptions from documents to do searches. Attributes should explain the document's properties and content. There are two types: basic type, which works as a library system and has a predefined set of attributes (e.g. author, title, ISBN), and intelligent type, which can extract information from a semantic network. One important issue is that there is no reliable way to find relevant documents without having to go through each one manually. As a result, the WWW Consortium has introduced RDF as a standard language for machine-readable descriptions of Web resources.

Automatic metadata generation - Automatic metadata generation can be used to collect context sensitive metadata, which is subsequently represented using RDF, according to Dewey decimal classification. Automatic classifier is a Java-based object-oriented system that gets HTML documents from a URL, analyzes the content, and applies a DDC classification class mark to the document. It compares terms discovered in the document to terms manually designated as DDC hierarchynodes. As a result, relevant metadata including the document title, keywords, abstract, and word count are generated. Documents with similar subject matter will be grouped together under the same classification mark. RDF is viewed as a "Web of Trust," with each document being well-described and generally recognized. Metadata has been attempted in HTML texts in a variety of ways. The issue is that such a technique is not required.

3.4 Information Retrieval Models

According to Lancaster and Warner (1992), the major information retrieval models are divided into four namely; Boolean model; Vector space model, otherwise called statistical model; probabilistic model; linguistic and knowledge-based models. These models were developed in order to retrieve information seamlessly.

a. Boolean Model

The Boolean model is a model used for current large-scale retrieval systems and in on-line information services. The model, which is based on a set of theories, comprised Boolean algebra and the Boolean logic components (such as ‘AND’, ‘OR’ and ‘NOT’. While the model is very useful in information retrieval, it has a major drawback in that it does not rank result list of retrieved information.

b. Vector Space Model (or Statistical Model)

The vector space model, sometimes referred to as the statistical model, is a type of information retrieval model that ranks documents based on the similarities between the search terms or query and the available documents. The model assigns high value to documents that have high number of the query terms in the document. So a document can have a few of the query terms in the title but may be ranked higher than other documents that have high number of the query terms in the title. The vector space model builds index by perusing through the documents to identify important terms in the documents. It is through this process that the document identifies important or relevant documents. Consequently, the vector space model thrives on two main assumptions:

1) The more similar a document vector is to a query vector; the more likely it is that the document is relevant to that query.

2) The words used to define the dimensions of the space are orthogonal or independent.

c. Probabilistic Model

According to Belkin and Croft (1992), the probabilistic model of information retrieval operates on the principle of ranking documents based on the probability of the documents being relevant to the researcher’s query or search terms. The model ranks documents by comparing the documents to the user query(ies). This is denoted by binary vectors ~d and ~q. One of the merits of this model is that it has higher chances of providing users with a ranking of high number of relevant documents. This is beside the fact that queries are easy to formulate in this model as users send queries using natural language and not query language. However, the model does not support the Boolean relation as it has a limited expressive power.

d. Linguistic and knowledge-based models

The linguistic and knowledge-based model is used to retrieve documents. The model does this by performing a morphological, syntactic and semantic analysis of available documents to determine the relevance of the documents to retrieve. This model is seen by many as the simplest form of document retrieval because the model uses search terms or keywords entered to search the documents keywords in the database. As a result, the model only retrieves documents by picking documents (in the database) with similar keywords entered by the user. This method of information retrieval has often been criticized since there is a good chance of the system missing some relevant documents. One major constraint facing the linguistic approach is it inability to resolve word ambiguities and/or generate relevant synonyms or quasi-synonyms based on the semantic relationships between words. However, the presence of advanced technologies such as experts systems and other forms of artificial intelligence has helped in the retrieval of relevant documents based on the linguistic and knowledge-based model.

4. CONCEPT OF MULTIMEDIA

The ability to gather, store, edit, integrate, and query information given in several formats, such as text, graphics, audio, video, and photographs, is referred to as multimedia. Multimedia isn't a technology in the traditional sense. It's more of a notion that represents a collection of technologies that work together to benefit the end user.

Multimedia is the field of computing which deals with the integration of text, graphics, videos, animation, images and any other type of media files for processing, transmission and storage (Ftsm, 2018). Basically, multimedia connotes that computer information can be represented through different media such as text, images, videos etc. It uses multiple forms of information content and information processing. The basic elements of multimedia include videos, images, text, audio, animation and graphics. Examples of multimedia include audio slideshow and video podcast. Contrary to traditional media like printed materials, multimedia combines information in diverse kinds of formats. Multimedia can be used to store information and its contents viewed through technologies like laptop, smart phones, and smart television. Moreover, multimedia can be broadly categorized into two broad categories, namely:

1. Linear category: In this category, there is smooth progression of content and it does not require any form navigation control from the viewer to make progress. An example of this is when you watch a blockbuster movie.

2. Non-Linear category: This category makes progression through the interactivity of the participants involved. For instance, the use of video games for entertainment requires the direct involvement of the actors.

4.1 MULTIMEDIA INFORMATION RETRIEVAL SYSTEMS

This refers to the technology used in the searching for and finding of multimedia documents (usually in the form of text, video, images, music etc) stored in a database or in the internet. According to Kambau and Hasibuan (2017), multimedia information retrieval system is an information retrieval system which seeks to extract information from multimedia data sources. This retrieval system allows one to retrieve and utilize text, images, audio and video after a search process. It is imperative to add that while multimedia information retrieval system refers to the technology or software used in the extraction of multimedia contents, multimedia information retrieval refers to the process of searching and retrieving multimedia documents. The retrieval of multimedia contents typically begins with a query (which could be in the form of text or an image) for the purpose of retrieving similar contents.

4.2 Content-based Multimedia Information Retrieval

Content-based multimedia information retrieval is used to retrieve image, audio and video content. Hence, it is divided into three categories namely; content-based image retrieval, content-based audio retrieval and content-based video retrieval. The content-based image retrieval makes use of elements such as colour, shapes and texture to retrieve image content from the database while the content-based video retrieval is an extension of the image retrieval but applies motion features to the moving image. The video retrieval system uses video parsing, content analysis features and abstraction to retrieve relevant videos from the database. The audio retrieval system makes use of audio signals with acoustic or semantic features. In this case, acoustic feature could include the pitch, cestrum and loudness while semantic feature could include rhythm, timbre and events.

4.3 Context-based Multimedia Information Retrieval

This type of retrieval combines the user's context and knowledge of information retrieval with multimedia search technologies to retrieve needed contents. The implication of this is that the user search behaviour and characteristics will influence the search outcome. Therefore, context-based multimedia information retrieval may include several dimensions such as time, location, user, current task.

4.4 Concept-based Multimedia Information Retrieval

This category of multimedia information retrieval focus on using manually built thesauri or by extraction of latent word relationship and concept from the corpus to retrieve relevant multimedia contents. It uses a classifier to select multimedia data from a pool of data or from a database when a query is made. It became necessary to develop the concept-based multimedia information retrieval due the inability of the content-based and context-based multimedia information retrieval to describe semantic visual features or semantic audio features.

4.5 Basic Multimedia Search Technologies

Among the basic multimedia search technologies or retrieval method are the metadata driven search retrieval, piggy-back text search, automated image annotation, fingerprinting amongst others.

- **Metadata Driven Search**

  The metadata provide information that is used to describe, manage, store, create and retrieve multimedia information. Metadata could be used in multimedia search queries and it comes in layers which enable one to submit a piece of information about an image, video, audio or text in a database. Multimedia metadata also play a major role in the indexing, classification and location of multimedia contents through the development of appropriate schemas. For example, the World Wide Web consortium uses the XML schema language to document and retrieve contents. Furthermore, several standards for bibliographic metadata have been developed for easy management of stored documents. For instance, the Machine-ReadableCatalogue 21 (MARC 21), MPEG-7 and the Dublin Core are among the most used metadata standards.

- **Piggy-Back Text Search**

  Simply put, this kind of multimedia search allows for automated processes to create text surrogates for multimedia. This category of multimedia search focus on the retrieval of multimedia contents, especially videos by extracting its search strings from Teledex, closed caption and subtitles. While the speech recognition and optical character recognition is used for audio and text retrieval respectively. However, information about some audio sounds like music could be converted into text like the MIDI files that also has some note representation. The piggy-back text search employs a variety of text matching techniques to search, edit and manage multimedia contents.

- **Automated Image Annotation**

  This is an image retrieval system where metadata are automatically assigned to digital image by computer systems using keywords or capturing to locate, organize and retrieve relevant images from a database. This process is sometimes referred to as automatic image tagging and can be likened to a multi-class image classification with a very large number of classes. The automated image annotation extract requested images by conducting an image analysis through machine learning technology. This retrieval system holds more advantages than the regular content-based image retrieval as queries can be more naturally specified by the user. Queries do not have to rely heavily on colour and texture as it is with the content-based image search.

- **Fingerprinting**

  This is an important retrieval method in multimedia information retrieval. This retrieval method uses a prescribed algorithm to identify and retrieve a specific multimedia object based solely on its content. Fingerprints can be applied to both audio and image. In fact, both set of media has the same fingerprints requirements, i.e., they are small, reliable, fast, and have unique location of different records in the database even under degradation or little change. However, the primary distinction between audio and image fingerprinting is that while the audio is a one-dimensional air pressure function of time, the image fingerprint contains static two-dimensional colour distributions (Rüger, 2010). The fingerprint retrieval system is an ideal system against degradation or deliberate modification. Some of the techniques involved in multimedia fingerprinting include computing of salient points, extracting features from vicinity, making invariant under rotation, quantizing, indexing or text search engines, enforcing of spatial constraints after retrieval etc.

5. INFORMATION RETRIEVAL ON THE WORLD WIDE WEB

The World Wide Web is an interconnected information system which gives universal access to wide range of documents with a uniform resource identifier. The Web contains web resources which are hyperlinked by means of hypertext or hypermedia and the linked documents connect related information which allows users to access them directly by clicking on the linked word or phrase. Being an information system, the web contains information resources in different media which include images, audio, video, animations, text and other forms of information usually with hypermedia feature. Invented in 1989 by Sir Tim Berners-Lee, the web operates with some software applications and standard protocols like the Hyper-Text Transfer Protocol (HTTP), Hypertext Mark-up Language (HTML) and Web browser. The hypertext mark-up language for instance, is used to format Web pages which usually contains resources with common theme and domain name. The HTTP on the other hand, is an internet protocol that allows the retrieval and transfer of linked resources across the web. These web technologies aid the retrieval and transmission of information over the internet.

5.1 Search Engines

The search engine is a web-based tool or software system that is used to locate and retrieve information on the web. They are primarily used to conduct web searches in a systematic way using the supplied web search query. Examples of search engines include Google, Microsoft Bing, Yahoo, Ask.com, Duck Duck Go, Alta Vista, Lycos etc. However, it is important to note that all of the above examples of search engines fall under the general-purpose search engines while search engines like mamma.com, dogpile.com and Meta Crawler can be categorized as Meta search engines. Meta search engines collect results or documents from several search engines in order to provide relevant documents. In summary, Meta search engines ranks results from other search engines.

5.2 Classification of Web Search Tools

Web search tools aid the retrieval of information from the World Wide Web. These tools use robots to index documents in the web before presenting a result via an interface. The search tool house different components and could be classified into type 1 search tool and type 2 search tools (Gudivada, Raghavan, Grosky & Kasanagottu, 1999). Gudivada et al (1999) distinguished between the type 1 and type 2 search tools using different dimensions such as the indexing techniques, strategies for query-document matching, methods of web navigation, query language or specification scheme amongst others. One major distinction between the type 1 and type 2 search tools is that while the type 1 search tools completely hide the organization and content of the index from the user, types 2 do not. In fact, a well-known characteristics of type 2 search tools is that its hierarchically organize subject catalogue or directory of the web and makes it visible to end-users as they search (Gudivada, et al; 1999). Examples of the type 1 search tools include Alta Vista, HotBot, Excite and Lycos while examples of the type 2 search tools include Magellan, Yahoo, and WWW Virtual Library.

5.3 Web Search Services

The web search services include activities by computer programmes to extend users' query from one search engine to others. The search services send out queries to multiple search engines and information sources simultaneously in order to retrieve comprehensive results. For instance, the MetaCrawler is an example of search services. When in action, the MetaCrawler broadcast user's queries to several separate search engines. Some of the search engines often include Infoseek, WebCrawler, OpenText, Lycos, Excite, Yahoo, AltaVista and Galaxy. Furthermore, aside broadcasting queries to other search engines, search services also merge results retrieved, exclude redundant5or duplicate information and present the final output as hypertext markup language page with clickable universal identifier. The search services primarily help to simplify web search by providing layer of abstraction to the user over several search tools.

5.4 Web Search Strategies

In searching for information on the web, there are two main classes namely; the known item search and the unknown item search. While the known item search is conducted when the information seeker knows the item s/he is looking for, the unknown item search is carried out when the information seeker is unaware about the availability of the information, he/she seek. For a known item search, the user may have some information such as the title of the work, the name of the author, the International Standard Book Number (ISBN) etc. However, in web searching, there are different approaches or search strategies to ensure information are easily retrieved from the web. Among the search strategies are:

a. **Keyword and Phrase Search**

   The keyword search strategy is the simplest and oldest form of web search strategy. This search method is done by entering a keyword or phrase in the system and the system then conduct an inverted file (index search) for each keyword. The keywords or phrase used in this search strategy is often retrieved from the subject heading list. The phrase search matches the phrase entered to other keywords in the system to retrieve relevant results. If the researcher has need to combine multiple keywords or phrases, then the researcher can introduce the Boolean operators or logic (AND, OR, NOT). To conduct a keyword or phrase search, the researcher must enter the keywords or phrases in the system using the keyboard or select the keywords from an index file or vocabulary control tools like subject headings list.

b. Boolean Search

This is the search technique that combines search terms or phrases using certain logic or operators known widely as Boolean logic. The Boolean logic helps to narrow or expand a search by introducing some basic operators such as AND, OR, NOT. While the AND operator allows the research to combine search terms of keywords and retrieve wider results, the NOT operator helps to narrow down the search results by eliminating related terms from the result. The `NOT` operator seek to retrieve only needed information while irrelevant information and `noise` are excluded from the search results.

c. Truncation

In truncation, a researcher is able to search for different forms of words by searching using the lexeme or root word. In this case, the researcher identifies a word having the same common root words and uses the root words to search for ords with similar root word. For instance, the word `astro` (relating to star or celestial objects) could be used to search for words such as astronaut, astronomy, and astrophysics. Similarly, the `acri` (relating to bitter) could be used to search for meaning of words such as acrid, acrimony, acridity etc. However, it should be noted that different forms of truncation exist. For instance, we have the right truncation, left truncation and making letters in the middle. The earlier examples of `astro` and `acri` are examples of right truncation. The left-turn truncation is used when one need to retrieve all words having the same characters at the right. Example includes words ending with `HYL` such as Methyl, Ethyl etc. Middle truncation retrieves words having the same characters at both left- and right-hand side. Irrespective of the type of truncation, it must be stated that poor use of truncation in the search and retrieval of information from the web could lead to truncation error.

d. Proximity Search

This type of search strategy is commonly used when searching for contents in CD ROM and in online databases. In proximity search, the user or information seeker is able to determine if two or more search terms should occur adjacent to one another; or if one or more words occur in between the search terms. In addition, the proximity search also enables the user to determine if the search term should occur in the same paragraph regardless of the intervening words. The user is at liberty to use different operators in proximity search and such operators used may differ from one system to another.

e. Field-specific Search

The field-specific search is primarily used for electronic databases were an information seeker can search for all fields in a database or may limit the search to a particular field or fields. A field in this context refers to a single piece of information from a record. It could also be seen as a data structure for a single piece of data. For example, in a table called students record, matriculation number, name of students, level of study etc. will each be a field. As such, field information will vary from one database to another.

f. Limiting Search

Limiting search occurs when a researcher narrows a search by introducing certain criteria such as the type of information source, the language, date, title or the year of publication. By using these parameters, the information seeker is able to limit or narrow down the search results. However, the choice of which parameter to use to limit the search result is heavily dependent system design or database concerned.

g. Range Search

The range search is primarily used to select record or data within a certain range. This type of search strategy is mainly used when retrieving numerical information. Some of the symbols used include the less than and greater than operators. Other operators are Greater than (>); Less than (<>); Greater than or equal to (>=); Less than or equal to (<=). The retrieval of information from the World Wide Web is gaining global attention even as more users are seeking for ways to easily retrieve relevant information from the Web, given the continuous exponential increase of information resources (including multimedia information) on the Web. This is coupled with the fact that there is continuous advancement in Web search tools. For instance, the introduction of Web browsers like Chrome has opened up newer functionalities for end-users and has increased their capability to effectively navigate through the Web. It is, however, envisaged that the current limitations associated with information retrieval on the web will be reduced, if not completely eliminated with the introduction of the semantic web which will enhance existing Web content with semantic structure in order to make it meaningful to computers as well as to humans.

6. CONCEPT OF DIGITAL LIBRARIES

The term digital libraries have been used interchangeably with other related terms such as online library, virtual library and digital repository. As a result, several scholars have given different explanation of what the term means. According to the Digital Library Federation (1998), the term digital library refers to "organizations that provide the resources, including the specialized staff to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities". However, a more generally accepted definition of digital libraries was given by David (2012). David defined digital libraries as a collection of documents in organized electronic form, available in digital format and which requires the use of digital technologies to appreciate its contents. The content in a digital library could range from digital objects such as text, videos, audio, images, magazines, books etc. Beyond aiding the storage of information resources, digital libraries are used to organize, search and retrieve information resources contained in the database.

According to Solvberg (2020), the primary goal of any Digital Library (DL) is to meet the demands of its users. The ability to access digital assets through a computer network is a must. The fact that Digital Libraries' information is maintained, permanent, and reliable is critical. Digital collections, a working environment, and technology and services make up a Digital Library. Information Discovery, or how to find information in the Internet environment, is a common issue. Digital Libraries are collections of digital items that include text, sound, maps, movies, photographs, and other media. It will be discussed how metadata may be used to describe digital items. Metadata is frequently categorized as follows: descriptive, structural, and administrative, and as a result, it supports more labor duties than only information retrieval (IR). The IFLA paradigm and the Resource Description Framework (RDF) are described, as well as other metadata formats (MARC, Dublin Core, and others) (Solvberg, 2020).

Digital libraries may be accessed online or offline, they may also vary in sizes depending on the storage medium used and the capacity of the medium. The scope of digital libraries could also vary from one library to another and it is mainly dependent on the category of users the library was established to serve. The information bearing materials in a digital library could include all disciplines and all areas of human knowledge. As opposed to the traditional libraries, digital libraries are not limited by geographical location (especially when resources can be accessed online) and time.

6.1 Features of Digital Libraries

Digital libraries have some distinct features which make it different from traditional libraries and give it an edge of the physical library. Cleveland (1998) identified the following asfeatures of digital libraries:

1. No Physical Boundary (Libraries without walls): One prominent feature of the digital library is that it is not confined by physical boundary. The content in the library could be accessed, updated and retrieved within any given location since the library is often designed within a given network such as the Local Area Network (LAN) Municipal/Metropolitan Area Network (MAN) and Wide Area Network (Internet) (WAN).

2. Require Digital Technology to Access its content: A unique feature of digital library is that it requires the users of the library to use digital technologies such as laptop or desktop computer, computer networks, Tablets, smart phones, etc. to access its contents. This is because resources in digital libraries are usually in digital form.

3. Multiple and Simultaneous Access: As opposed to the conventional library were only one user can use an information resource at a given time, the digital library allows for multiple or simultaneous access to information resources.

4. Space Requirement: The digital library is able to store large amount of information resources in a small space without taking too much physical space. This makes it possible for many traditional libraries to switch to hybrid libraries.

5. Preservation and Conservation: Information resources stored in digital libraries offers long term preservation to materials that would ordinarily suffer from degradation as a result of repeated use.

6. 24/7 or Round the Clock Access: One notable feature and advantage of the digital library is that it provides round the clock access to it resources. The implication of this is that library users are able to access the content of digital library at any given time. This gives it advantage over contents in traditional libraries that can only be accessed during library opening hours.

7. Easy Retrieval of Information Resources: A key feature of the digital library is that its contents can be easily retrieved. Many digital libraries provide friendly user interface which allows users to search for documents using search terms such as title, words/phrase, name of authors, subject etc., to search the entire collection.

6.2 Digital Libraries and Information Retrieval

The idea of digital libraries and information retrieval shares common objective of helping to meet the information needs of users through easy access to information. However, digital libraries, more often than not, houses diverse kinds of content such as bibliographic records, full text content, aggregated contents amongst others. Therefore, to understand information retrieval in digital libraries, it is useful to group the different types of knowledge-based information so as to comprehend issues of information retrieval in digital libraries.

6.2.1 Bibliographic Content/ Databases

The bibliographic content is often seen as the first category of electronic content in a digital library. This content primarily constitutes the bibliographic citation of information resources such as books, article, videos, images and other kinds of information materials. The citation could include the name of the author, title of the content, publisher's information, and keywords amongst others. In the case of a journal article, the citation could include the volume, issue and page number. The bibliographic content, commonly found in bibliographic databases of a digital library, has been the mainstay of information retrieval systems for a long time now. The content in this case does not contain the actual information being sought by a user but is a pointer to where the information can be accessed.

6.2.2 Full Text Content/ Database

As the name implies, this type of content or database contains full text of a publication. The full text search in a database is widely known as document search and it is a common search conducted in digital libraries. This type of database started to gain prominence from the 1990s when advancement in the technology made it economically and technologically possible for full content to be stored using technology. The content in this category consist primarily of E-books, journal articles, reports and other kinds of document. The full text database is considered an improvement to the bibliographic database as it gives users the actual content being sought for. Information retrieved from the bibliographic databases could be used to retrieve content from the full text databases. For instance, Research Library in GALILEO provides information on the citation and the entire content (text).

6.2.3 Aggregations

This is another category of knowledge-based information in a digital library in connection to information retrieval of content. In the context of digital library, aggregation loosely refers to the process of grouping or combining of two or more entities to more a new and more meaningful entity. Aggregation of content combines the different kinds of content such as citations and full text content to make a whole meaningful new content. This means that aggregations have a wide variety of different types of information serving diverse needs of information seekers or users. Aggregated contents are developed for all kinds of users, such as students, practicing professionals and researchers. Example includes the MEDLINE plus which has both full text content and bibliographic content (citations) aggregated for ease of access to a particular topic. Other examples of aggregators for academic databases are EBSCO, ProQuest, and Dialog etc.

6.3 Indexing in Digital Libraries

Digital libraries, just like the Internet and other repositories indexes its contents to make it more easily discoverable and accessible to users. Indexing in digital libraries makes search and retrieval of information resources faster than in conventional libraries. The indexing in this case could be done manually using controlled vocabulary or it could be in automated form. The manual indexing is usually done by human indexers who use standard terms, keywords and synonyms to assign indexes to an item or a document. The indexers follow a set of protocol which ensures that only terms that are the canonical representations of the content are used to generate the index. The automated indexing, on the other hand uses computer programs and technologies to generate indexes for content by using notable words in the document as index terms. This method of indexing is what is currently being applied to most content in digital libraries has proliferation of information resources (information explosion) has made it impracticable to manual indexing to be carried out on all published document.

6.4 Challenges Of Digital Libraries

Digital libraries face a lot of challenges which are highlighted below.

1. **Infrastructure constraints** – The main roadblock here is a lack of high-capacity bandwidth for network and internet access, as well as a weak computer infrastructure in libraries.

2. **Lack of professional expertise –** In today's changing climate, where technological developments push librarians to embrace and adapt to the changes or risk being bypassed on the information highway, there is a lack of professional expertise.

3. **Absence of high-quality contents —** Nigeria has a rich heritage of arts, folklore, spirituality, traditional knowledge, and so on, all of which are unused and may have been destroyed as a result of disasters, terrorism, and conflict. As a result, because information is largely in print, it is not easily transmitted or retrieved. In order to tackle the problem, they must be digitized in order to have access to or retrieve the extraordinary information.

4. **Lack of strong digital policy -** In Nigeria and other developing countries, most libraries lack ICT planning and strategy plans to address the issues brought by technological advancements, information overload, and user demand for effective retrieval. Legislators who are unaware of or unfamiliar with the requirements of digital preservation are more likely to pass laws that either neglect or inadequately address digital preservation challenges. As a result, digital resource preservation is not taken seriously.

5. **Technological Obsolescence -**Markets are full of a variety of digital formats that continually change from time to time with some formats getting obsolete (Caplan, 2004). Technological obsolescence comes as a result of continuous upgrade of operating systems, programming languages, applications and storage media. Such changes or updates make preservation of digital materials meaningless. For example, a Ph.D. dissertation submitted in the School of Postgraduate Studies was backed-up in a floppy disk before submission five years ago. The researcher after the five years, wanted to access the thesis so as to show it to his colleague working on related topic. Unfortunately, he was unable because the generation of computers currently available was never and could not open the disc due to technology updates or decay. The available hardware does not have the right drive for the researcher to access the information and therefore rendered the dissertation inaccessible (Asogwa, Ilo, Asadu, Igbo, & Asogwa, 2021).

6. **Copyright Legislation** -Most national legislations do not clearly express the copyright of software required to access digital material, as well as the right to copy for preservation. For example, due to copyright rules, a subscriber to an internet-based information service is obliged to renew the access license on a regular basis, even for items that have been paid for a long time, in order to continue viewing the same information. Another issue is that digital progress has been far too fast and expensive for governments and organizations to adopt timely and well-informed preservation policies.

7. **Lack of Collaboration and Partnership -** Another important issue with digital information is the absence of coordination and communication among libraries and stakeholders, as well as the lack of explicitly defined responsibility for the long-term preservation of digital assets. Governments, curators, publishers, relevant companies, and heritage organizations are all lacking in collaboration and partnership.

8. Lack of Disaster Preparedness- Another issue with digital preservation is the risk of digital media being lost in the event of a calamity such a fire, flood, device failure, or virus assault. In the absence of crisis preparedness, planning, and migration actions, precious information resources in libraries are frequently lost forever.